Linguistica dei corpora

SOCIETÀ DI LINGUISTICA ITALIANA

SLI 58**

LA LINGUISTICA ITALIANA ALL'ALBA DEL TERZO MILLENNIO (1997-2010)

BULZONI

ROMA 2013

LA LINGUISTICA ITALIANA
ALL'ALBA DEL TERZO MILLENNIO (1997-2010)

SLI 58**

BULZONI 2013

PUBBLICAZIONI DELLA
SOCIETÀ DI LINGUISTICA ITALIANA
58**

SLI
SOCIETÀ DI LINGUISTICA ITALIANA

LA LINGUISTICA ITALIANA ALL'ALBA DEL TERZO MILLENNIO (1997-2010)

a cura di
GABRIELE IANNÀCCARO

Tomo secondo

BULZONI EDITORE ROMA 2013

ISBN 978-88-7870-908-9

00185 Roma, via dei Liburni, 14
http://www.bulzoni.it
e-mail: bulzoni@bulzoni.it

INDICE

Tomo primo

Tomo secondo

PARTE SECONDA

LIVELLI DI ANALISI E MOMENTI DI RICERCA

MANUEL BARBERA
(*Università di Torino*)

Linguistica dei corpora[1]

1. INTRODUZIONE

La novità più cospicua di questo decennio è resa palese dall'esistenza medesima di questo capitolo: l'avvenuto sdoganamento della linguistica dei corpora (od all'inglese *Corpus Linguistics*) dalla più generale linguistica computazionale e dalle sue molteplici anime, che spaziano dal trattamento automatico delle lingue naturali (TAL o NLP, *Natural Language Processing*) all'intelligenza artificiale (IA), qui difatti oggetto di un differente capitolo. La maturazione di questa disciplina, posta ad un crocevia tra le tecnologie di TAL e le pratiche filologiche, lessicografiche e di storia della lingua (cfr. Barbera, 2011: 27), già ben nota e da tempo strutturata in àmbito linguistico anglofono, anche nella tradizione linguistica italiana è avvenuta non senza differenze teoriche significative (come si illustrerà nel § 2.1), radicate nelle diverse tradizioni. Se le applicazioni sono ormai svariate e vanno dal più tradizionale campo della lessicografia (cfr. qui il contributo di Carla Marello) a quelli della linguistica contrastiva o apprendologica, dalla morfologia, alla semantica ed alla linguistica testuale (un buon campionario è offerto dalla silloge di Barbera, Corino, Onesti, 2007), va però detto che i corpora di pubblico accesso non sono poi moltissimi (forse anche per le ragioni legali affrontate nel § 2.4).

Nella presente rassegna privilegeremo, da un lato, le riflessioni teoriche e storiografiche sulla disciplina tutta (intendendosi che le applicazioni di tale disciplina a specifici domini linguistici dovrebbero comunque risultare coperte da altre sezioni di questo volume), e dall'altro i corpora di lingua italiana effettivamente prodotti. Non ci prefiggiamo certo la completezza (comunque impossibile in questi spazi), ma piuttosto l'esemplarità, limitandoci a delineare i filoni e le opere che ci paiono storiograficamente centrali. I limiti (naturalmente applicati con la dovuta elasticità) saranno dettati dalla effettiva pertinenza dei prodotti ad una definizione stretta di corpus (che sarà introdotta e discussa nel § 2.2) e dalla loro reale e libera messa a disposizione pubblica (stante la centralità della questione legale discussa nel § 2.4), ad esclusione, quindi, di quanto sia eclusivamente proprietario o commerciale. Questa scelta implica necessariamente anche il ridimesionamento della sezione sul *software*, dove l'*open source* e la distribuzione libera costituiscono ancora una percentuale minoritaria, anche se non insignificante: cfr. § 4.

Porsi una limitazione alla produzione italiana è quasi inutile, dato che corpora italiani non commerciali prodotti all'estero praticamente non ve ne sono, e quindi la cosa

[1] Tutti i collegamenti web sono stati controllati nell'aprile 2012.

va praticamente da sé: oltre al materiale italiano mantenuto all'IMS Stuttgart ed alle sperimentazioni con TactWEB di corpora giornalistici di Elisabeth Burr, inizialmente pubbliche ma ormai da anni commercializzate (cfr. Burr, 2004 e pertanto comunque fuori da questa rassegna) vanno ricordate forse solo la componente italiana di CHILDES (cfr.§ 3.4) ed il corpus di italiano ticinese di Pandolfi, 2006 (cfr.§ 3.5).

Un'ultima avvertenza, che va anche a confermare quanto sopra si diceva a proposito del radicamento nella tradizione grammaticografica italiana e della intrinseca maturazione della disciplina, concerne il trattamento dei numerosi anglicismi tecnici che vi sono invalsi, che sono stati ripetutamente oggetto di studio e normalizzazione, prima da Carla Marello e da chi scrive (Barbera, Marello, 2000) in un importante convegno dell'Accademia della crusca, e poi da parte del sottoscritto (Barbera, 2007; Barbera, 2009, pp. 7-13, § 1.4); ed è alla proposta globale di Barbera, 2009, cit. con tutte le conseguenze pratiche che implica («Tondo (invariabile) o corsivo (con plurale in *-s*)? Prestito non adattato (ma comunque accettato, fosse anche *faute de mieux*) o fastidioso termine straniero se non da puristicamente evitare almeno da porre nella quarantena del corsivo?», Barbera, Marello, 2009, n. 3), che ci atteniamo nel presente contributo.

2. TEORIA E STORIOGRAFIA

2.1. *Procedimento "corpus based" ed empirismo all'italiana*

La linguistica dei corpora anglosassone (anzi, *corpus linguistics*) si è di solito voluta presentare come una radicale novità, accentuando gli aspetti quantitativi sui qualitativi, e contrapponendosi, a volte in modo esasperato, al generativismo come roccaforte empiristica, perlopiù in modo assai generico (cfr. il classico ed emblematico manuale di McEnery, Wilson, 2001, § 1) e più raramente in modo meditato e filosoficamente consapevole (Sampson, 2001); così l'enfasi è vertita sul ricorso esclusivo ai dati presenti nei corpora, spesso ipostatizzati come soli oggetti linguistici possibili (il cosiddetto procedimento *corpus driven*: cfr. Sinclair, 1991) in palese ostilità all'introspezione (cfr. invece Renzi, 2008) propugnata dal paradigma generativo. In Barbera, 1909 e 2011 (ed ora 2012, in stampa), si è invece tracciata una storia di sostanziale continuità con la tradizione della linguistica filologica otto-novecentesca, riconnettendosi al filone, meno integralista, cosiddetto *corpus based*, teoricamente definito (ma non inventato) a partire da un fondamentale contributo di Fillmore, 1992: dai fatti di *parole* raccolti in un corpus si può risalire ai loro correlati stati di *langue* (contro i generativisti più ortodossi, Chomky *lui même* compreso: cfr. da ultimo Andor, 2004), anche se certamente non tutti gli elementi di una *langue* saranno contenuti in un corpus (contro i più accesi antigenerativisti sostenitori della pratica *corpus driven*). È l'uso (testimoniato dai corpora), anzi, che fonda la *langue*, anche se i corpora essendo per definizione finiti (cfr. *infra*, § 2.2) ne rappresenteranno solo un sottoinsieme, significativo quanto più il corpus sarà stato costruito in modo accorto (gioco nel quale non può non rientrare la famosa introspezione): ciò, naturalmente, all'insegna della migliore tradizione wittgensteiniana (cfr. Barbera, Marello, 2009). Il dialogo con i generativisti meno intransigenti è così aperto, come dimostrano i rapporti tra le due imprese del Corpus Taurinense (Barbera, 2009) e di ItalAnt (Renzi, Salvi, 2010) e l'importante mossa di apertura di Renzi, 2008. Questa minore conflittualità ed apertura al dialogo (in cui probabilmente Renzi e chi scrive hanno avuto una discreta parte) è precipua caratteristica della situazione italiana, e sarebbe impensabile nelle aree anglofone.

Ma non solo. Francesco Sabatini ha ripetutamente argomentato (a partire da Sabatini, 2006) che il procedimento *corpus based*, l'idea che la norma si ricavi dall'uso, sta alla base della storia linguistica italiana stessa, visto che il Dizionario della Crusca, che di quella tradizione rappresenta un momento fondante, è proprio stato costruito su testi[2].

Ed è proprio la ritrovata dimensione testuale (perlopiù persa nella *corpus linguistics* di lingua inglese[3] a favore della puramente lessicografica, o al più variazionale) che è uno dei fenomeni nuovi e più vistosi della linguistica dei corpora italiana degli ultimi anni, da Barbera, Corino, Onesti, 2007 a Ferrari, Lala, 2010, certo complice la piena fruizione dei contesti, resa possibile dalla soluzione dei problemi legali di copyright (cfr. *infra*, § 2.4).

Una riflessione sull'àmbito, più tradizionale per la linguistica dei corpora, della statistica è invece stata utilmente portata avanti, in inglese ed in sede internazionale, da uno dei protagonisti principali della scena italiana, Marco Baroni (Baroni, Evert, 2009).

2.2. *La definizione specialistica di "corpus"*

Abbiamo più volte fatto riferimento ad una definizione tecnica e stretta dell'oggetto principe della linguistica dei corpora, che è un meditato risultato dell'ampia rassegna del 2007:

> Raccolta di testi (scritti, orali o multimediali) o parti di essi in numero finito in formato elettronico trattati in modo uniforme (ossia tokenizzati ed addizionati di markup adeguato) così da essere gestibili ed interrogabili informaticamente; se (come spesso) le finalità sono linguistiche (descrizione di lingue naturali o loro varietà), i testi sono perlopiù scelti in modo da essere autentici e rappresentativi. (Barbera, Corino, Onesti, 2007 a, p. 70)

[2] E la paternità della *moderna prattica* va ascritta proprio ad un italiano, il padre Roberto Busa S.J., con la sua pionieristica opera su Tommaso d'Aquino, come già argomentava Marello, 1996, pp. 167-8. Ed anzi, se il suo primo *Saggio* era nientemeno che del 1951, i risultati finali di più di mezzo secolo di alacre lavoro, cioè il fondamentale *Index Thomisticus* (http://www.corpusthomisticum.org/it/), sono stati pubblicati e messi liberamente a disposizione online nel 2005, rientrando quindi (anche se non veramente di corpus nei sensi del § 2.2 si tratta) nei limiti cronologici di questa rassegna.

[3] Alle importanti aperture di Douglas Biber su "generi testuali" e variazione testuale (cfr. almeno Biber, Finegan 1991 e Biber 1992) non ha infatti fatto séguito una robusta tradizione se non spostandosi sul versante strettamentre variazionale: i risultati "testuali" più rilevanti (cfr. ad es. Ramsay, 2003) riguardano piuttosto la linguistica computazionale che non quella dei corpora in senso proprio qui tematizzata.

Cui aggiungerei la seguente postilla:

> *Linguisticamente*, inoltre, un corpus è una raccolta di atti di *parole*, e dai fatti di *parole* raccolti in un corpus si può risalire ai loro correlati stati di *langue*, anche se certamente non tutti gli elementi di una *langue* saranno contenuti in un corpus: è l'uso testimoniato dai corpora, anzi, che fonda la *langue*, anche se i corpora essendo per definizione finiti ne rappresenteranno solo un sottoinsieme. (adatt. da Barbera 2012, *à paraitre*, e cfr. *supra*).

Una definizione tale, come evidenziato nella rassegna medesima, non si ritrova in genere nella letteratura internazionale, ed è un ulteriore segno del rigore della tradizione italiana.

2.3. *Le caratteristiche "tecniche": token, markup, tagging*

La definizione citata fa esplicito riferimento ad alcuni concetti irriducibili quanto spesso trascurati nella trattatistica: token e type, la cui definizione risale nientemeno che a Peirce (Peirce, 1933 [1906], p. 537):

> A common mode of estimating the amount of matter in a MS. or printed book is to count the number of words. There will ordinarily be about twenty *the's* on a page, and of course they count as twenty words. In another sense of the word "word", however, there is but one word "the" in the English language; and it is impossible that this word should lie visibly on a page or be heard in any voice, for the reason that it is not a Single thing or Single event. It does not exist; it only determines things that do exist. Such a definitely significant Form, I propose to term a *Type*. A Single event which happens once and whose identity is limited to that one happening or a Single object or thing which is in some single place at any one instant of time, such event or thing being significant only as occurring just when and where it does, such as this or that word on a single line of a single page of a single copy of a book, I will venture to call a *Token*. An indefinite significant character *such* as a tone of voice can neither be called a Type nor a Token. I propose to call such a Sign a *Tone*. In order that a Type may be used, it has to be embodied in a Token which shall be a sign of the Type, and thereby of the object the Type signifies. I propose to call such a Token of a Type an *Instance* of the Type.

Il mantenimento di tale distinzione, terminologica[4] e concettuale, come essenziale caratteristica di un corpus consente di ancorare la disciplina non solo alla statistica in generale (dove la percentuale di token e type è uno dei calcoli di base) ma anche alla migliore tradizione semiotica, logica e filosofica, all'insegna di quella sintesi di elementi matematici e linguistici che è caratteristica precipua della linguistica dei corpora (non a caso si è spesso parlato di "informatica umanistica" come suo iperonimo).

La nozione di markup, che segna il difficile confine tra testo e metadata, e quindi l'organizzazione del corpus, è altro concetto essenziale, non a caso centrale anche nella codificazione della internazionale Text Encoding Initiative (cfr. le *TEI Guidelines*, giunte ormai alla quinta versione: Burnard, Bauman, 2007), dove è stata elaborata da un illuminato logico e storico della filosofia italiano, Dino Buzzetti (cfr. Buzzetti, 1999).

Distinto da markup va considerato anche il tagging, che pure di esso è propriamente una forma, cioè l'associazione ad ogni token di specifici attributi informativi; il caso più tipico è quello della annotazione morfosintattica o PoS-Tagging, su cui, per l'italiano ed in Italia, si è molto lavorato. Già nel decennio precedente si era avuto l'importante precedente di Monachini, 1996 a tracciare le linee della questione, che è stata, variamente ed in modi diversi, ripresa ed elaborata da chi scrive (Barbera 2007a e b) e da Mario Baroni (Baroni, Zanchetta 2005), appoggiandosi al medesimo software (il TreeTagger sviluppato dall'IMS Stuttgart: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html) oltre che da Fabio Tamburini (Tamburini, 2007), con strumentazione propria. I tre progetti muovono in diverse direzioni: Barbera ha curato piuttosto l'architetura logica delle strutture tipate (cfr. Carpenter, 1992) e la prospettiva linguistica storico-teorica (Barbera, 2009) laddove il tagset di Baroni (http://sslmit.unibo.it/~baroni/collocazioni/itwac.tagset.txt) è orientato più all'efficacia computazionale che alla granularità linguistica e quello di Tamburini è ricavato in modo computazionalmente assai interessante dai dati medesimi (Bernardi *et alii*, 2006).

2.4. *Il "problema" legale*

Può sembrare strano che aspetti legali occupino una posizione di rilievo in questa rassegna, ma sottrarre i corpora dal limbo giuridico (software o opere a stampa?) in cui si trovavano è equivalso a sdoganarli dall'incubo del copyright, riallineando la linguistica dei corpora al più vasto movimento dell'open source, così facilitando la circolazione di risultati e risorse. Il problema (definito in Allora, Barbera, 2007 e Zanni, 2007) era, peraltro, assai sentito anche dalla comunità internazionale (cfr. le discussioni, nei primi anni del Duemila anche molto accese e sconfortate, apparse sulla *mailing list* Cor-

[4] Vi sono taluni che hanno invece preferito usare la coppia terminologica "occorrenza *vs* forma", rinunciando ai benefici dell'internazionalismo e della multidisciplinarità, ma soprattutto rischiando di creare quell'illusione che i type siano solo la mera classe dei loro token contro cui aveva così efficacemente messo in guardia Quine, 1987, pp. 216-9: le classi, infatti, devono essere oggetti completamente astratti, mentre le "classi di token" non lo sarebbero abbastanza per i type, con tutte le aporie che l'uso improrio dell'insieme vuoto notoriamente comporta. Inoltre «It is seldom appreciated that *occurence* is a third thing: not token, but something between. The word *der* has two occurrences in the sentence *Es ist der Geist der sich den Körper baut*; and I speak now of types, not tokens. Tokens occurr in tokens, types in types» (Quine, 1987, p. 218; e cfr. sopra il concetto peirceiano di *Instance of the Type*).

pora: http://www.hit.uib.no/corpora/), ma ne mancava una appropriata soluzione giuridica che è stata data solo in Ciurcina, Ricolfi, 2007, che presenta anche dei pratici modelli contrattuali. Basata su Creative Commons (http://creativecommons.org/ o in versione localizzata http://www.creativecommons.it/), e precisamente sulle licenze *Share Alike* (o *Condividi allo stesso modo*), si tratta di una soluzione italiana ma facilmente esportabile anche all'estero in quanto fondata su schemi internazionali.

Non è un caso che tanto il corpus PAISÀ (cfr. *infra* § 3.2) diretto da Marco Baroni quanto tutti i corpora diretti da Manuel Barbera (cfr. *infra* § 3.3) siano legati, sia pure variamente, a Creative Commons. Che è una riprova di come la presenza di corpora di completa disponibilità pubblica nel panorama italiano sia strettamente connessa ad una soluzione del problema legale: la democratizzazione della linguistica dei corpora deve necessariamente passare attraverso una rinuncia della logica proprietaria, spesso accademica, che attraverso un'interpretazione tradizionale della legge sui diritti d'autore porta solo a secretare i risultati della ricerca.

L'altra ricaduta è stata la consistente appropriazione della linguistica dei corpora da parte di quella testuale (come si diceva *supra* al fondo di § 2.1), fenomeno che per ora è solo italiano, propagato soprattutto dai gruppi di ricerca di Torino e di Basilea (svizzeri ma italofoni ed italianisti). La illimitata e piena fruibilità dei contesti (fino ai testi interi) è stata senz'altro un elemento determinante in ciò: la difettosa acquisizione dei diritti ha in passato (ed a volte anche nel presente come nel caso del CORIS, cfr. *infra*, § 3.1) portato a cautelative (ma legalmente spesso ingiustificate) restrizioni dei contesti ottenibili in pubblico (quando non a complete secretazioni dei dati).

3. LA PRATICA

3.1 *Tipologie e disseminazione*

I corpora di italiano prodotti nell'ultimo decennio coprono ormai tutte le principali varietà diamesiche ed alcune delle diacroniche della lingua: si va, per un corno, dallo scritto, al parlato ed alle più diverse forme dei media (italiano degli SMS, dei blog, di Usenet, trasmesso, ecc.), e per l'altro, dalla lingua contemporanea all'italiano del Duecento. Non tutti i corpora però sono facilmente e gratuitamente accessibili. Il che giustifica che la presente rassegna (che dell'accessibilità ha fatto un requisito determinante per l'inclusione) sia in larga misura per centri di produzione (fondamentalmente due, quello imperniato su Marco Baroni, § 3.2, e quello su Manuel Barbera, § 3.3; cui si aggiungeranno, raccolti in un terzo paragrafo, i residui, § 3.4) piuttosto che per tipologia di corpora. Questo per i corpora, sia pure in differenti declinazioni, di lingua scritta; per quelli di lingua orale il discorso è in parte diverso e più sfrangiato e richiederà un ulteriore paragrafo (§ 3.5) a sé stante.

Un ulteriore discorso a parte va inoltre fatto per il bolognese CORIS (*CORpus di Italiano Scritto*: http://corpora.dslo.unibo.it/TCORIS/), cui già abbiamo accennato, e per le altre analoghe iniziative del gruppo bolognese, come il BoLC (*Bononia Legal Corpus*: http://corpora.dslo.unibo.it/bolc_eng.html). Il CORIS è spesso citato come il più

importante strumento di riferimento per lo scritto italiano contemporaneo, il che non è propriamente vero: pur essendo eccellentemente costruito[5], il minimo che si possa dire, infatti, è che è scarsamente fruibile, dato che il suo accesso online, ora liberalizzato ma fino a poco tempo fa subordinato ad una complessa e scoraggiante anche se gratuita procedura di registrazione, è limitato a 300 risultati e restituisce indici KWICK (cfr. Manning, Schütze, 1999, § 1.4.5, pp. 31-34) con soli 30 caratteri di contesto per parte; e lo stesso vale anche per il BoLC. Ma per fortuna ci sono anche altre risorse.

3.2. *Da Forlì a Trento: l'officina di Marco Baroni*

Marco Baroni (http://clic.cimec.unitn.it/marco/research.html) ed il suo gruppo di ricerca ancora a Forlì aveva prodotto e reso liberamente consultabile online il corpus La Repubblica, costruito a partire da 16 annate dell'omonimo quotidiano. Indicizzato accuratamente col CWB (cfr. *infra*, § 4), di cui la maschera di interrogazione ben conserva la duttilità, con i suoi 326.363.463 token (dati da Baroni *et alii*, 2009) è già di dimensioni assai notevoli: più di 3 volte del BNC, *British National Corpus*, che, con i suoi 96.868.603, già costituiva un vero traguardo internazionale.

Il progetto WaCky (*Web-as-Corpus Kool Ynitiative: http://wacky.sslmit.unibo.it/doku.php*; cfr. Baroni, Bernardini, 2006 e Baroni *et alii*, 2009) è mirato alla costruzione di ancora più grandi corpora (di italiano, inglese e tedesco) a partire dal Web, seguendo una tendenza particolarmente attuale nella linguistica dei corpora (almeno a partire da Kilgarriff, Grefenstette, 2003), ma evitando la problematica infrazione alla regola della finitezza del corpus (cfr. Barbera, Corino, Onesti, 2007, pp. 44-45, § 1.5). Un gigantesco (1.585.620.279 token!) corpus italiano così allestito, itWaC, è (sia pure con qualche problema di copyright) già liberamente ottenibile, anche se non ne è ancora pronta un'interfaccia web (e le risorse server necessarie a gestire una simile mole di dati certo vi avranno parte).

Anche PAISÀ (*Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati*: http://www.corpusitaliano.it/it/), è costituito da testi tratti dal web (raccolti nel settembre/ottobre del 2010; per i criteri cfr. Borghetti, Castagnoli, Brunello, 2011), ma questa volta completamente cautelandosi dal punto di vista legale, in quanto sono accolti solo testi licenziati sotto *Creative Commons Share Alike*: libero pertanto da diritti di sorta, conformemente a quanto teorizzato in Ciurcina, Ricolfi, 2997 (cfr. qui § 2.2 *supra*), il corpus è tanto scaricabile quanto agevolmente consultabile online. Di ampia ma non "esagerata" dimensione come itWac (si tratta pur sempre di circa 250 milioni di token) e completamente annotato, trascende ampiamente le finalità glottodidattiche per cui è nato.

[5] Tecnicamente è pur sempre un prodotto della eccellente mano di Fabio Tamburini (http://dslo.unibo.it/People/Tamburini/), uno dei migliori linguisti computazionali presenti attualmente sulla scena italiana.

3.3. *Torino: l'officina di bmanuel.org*

Capitanato, invece, da Manuel Barbera (http://www.bmanuel.org/personal/barbera/barbera.html), centrato su bmanuel.org (http://www.bmanuel.org/), appoggiato all'ormai fu Dottorato in ingegneria linguistica ed al sito di distribuzione dell'Università di Torino (http://www.corpora.unito.it), e forte dell'originaria iniziativa di Carla Marello, è questo il gruppo che in Italia ha sperimentato col CWB (cfr. *infra*, § 4) da più lunga data (almeno dal 1998), producendo e mettendo in libera distribuzione (requisito che abbiamo sempre posto come determinante: tutte le risorse sottoelencate sono consultabili online usando la regolare sintassi di interrogazione del CWB, cfr. *infra* § 4) un certo numero di risorse, e muovendosi su diverse tipologie di corpora.

Il CT (*Corpus Taurinense*: http://www.bmanuel.org/projects/ct-HOME.html) è un corpus storico (la categoria è spesso, meno accuratamente, riferita come "corpora diacronici") di italiano antico[6], ormai giunto alla sua seconda ampliata release (CT+ o neo-CT: disponibile alla medesima homepage). Di modeste dimensioni (attualmente 270.872 token nel CT+) ma accuratamente e riccamente annotato oltre che ampiamente documentato (CT: Barbera, 2008; CT+: Barbera, 2012), rappresenta la punta di diamante della sperimentazione di Manuel Barbera che, con la collaborazione determinante di Marco Tomatis (http://www.bmanuel.org/personal/tomatis/tomatis.html), nonché di molti altri, è il responsabile del progetto.

I NUNC (*Newsgroups UseNet Corpora*: http://www.bmanuel.org/projects/ng-HOME.html) sono una innovativa suite di corpora multilingui (ma l'italiano vi ha avuto sviluppo privilegiato: cfr. Barbera, Marello, 2011) basati sui testi delle gerarchie nazionali di Usenet (cfr. Barbera, 2011), dal 2003 ad oggi; il progetto, di cui sono già stati pubblicati cospicui risultati, è tutt'ora in corso.

L'Athenaeum Corpus (http://www.bmanuel.org/projects/at-HOME.html) è un piccolo corpus di prosa accademica prodotta nell'Università di Torino.

Attualmente in corso vi sono ancora Jus Jurium (http://www.bmanuel.org/projects/ju-HOME.html), un corpus che vorrebbe documentare il discorso giuridico esistente in Italia in tutti i suoi generi, con speciale attenzione agli aspetti testuali e diplomatici (cfr. Onesti 2011), ed il Corpus Segusinum (http://www.bmanuel.org/projects/vs-HOME.html), di cui è già disponibile una beta online, che è il primo di una suite di corpora tesi ad esplorare le testate della stampa regionale (cfr. Barbera, Onesti, 2010).

In corso, ma già interrogabili, segnalo anche gli SMS Monitor Studies (http://www.e-allora.net/SMS/ms_index.php), un corpus di SMS al momento di soli 1.394 messaggi, ma in crescita, mantenuto da Adriano Allora (http://www.e-allora.net/).

Un discorso a parte va fatto per il corpus di italiano L2/LS VALICO (*Varietà di Apprendimento della Lingua Italiana Corpus Online*) e per il suo corpus di controllo L1 VINCA (*Varietà di Italiano di Nativi Corpus Appaiato*) recentemente migrati su un dominio indipendente (http://www.valico.org/) ed ora ad esclusiva cura di Carla Marello

[6] Nel senso, affatto renziano, di fiorentino del secondo Duecento.

ed Elisa Corino (cfr. Corino, Marello, 2009). Di interesse apprendologico e didattico esclusivo ma spiccato, presentano una grande cura ed abbondanza soprattutto nel trattamento dei metadata sociolinguistici (cfr. Allora, Colombo, Marello, 2011).

3.4. *Altri centri e tipologie*

Se le fila principali della linguistica dei corpora italiana, almeno nella sua accezione "free", si dipanano tra questi centri, non mancano altre voci, spesso devolute a tipologie speciali.

Un centro di eccellenza molto importante ma di solito legato a logiche proprietarie è l'EURAC (*EURopean ACademy of bozen/bolzano*). Tuttavia, il suo corpus LexAlp, costruito e gestito col CWB (cfr. *infra*, § 4), che mira ad un «raffronto contrastivo tra i linguaggi giuridici utilizzati dagli stati dell'arco alpino, con la successiva armonizzazione dei termini principali per la comunicazione sovranazionale», secondo recita la homepage, è invece liberamente consultabile online; le lingue coperte sono francese, italiano, tedesco e sloveno.

Oltre ai corpora multilingue, un'altra tipologia speciale è quella dei cosiddetti "corpora diacronici" (*recte* "storici"), di cui abbiamo già parlato a proposito del Corpus Taurinense (cfr. *supra*, § 3.3), e di cui vantiamo ormai una ricca tradizione, ma purtroppo solo per la fase antica della lingua italiana. Innanzitutto va menzionata la banca dati dell'OVI (*Opera del Vocabolario Italiano*; cfr. anche *infra*, § 4), un grandioso e fondamentale database testuale di italiano antico (http://tlioweb.ovi.cnr.it); liberamente consultabile, mantenuto dall'OVI (http://www.vocabolario.org/) e diretto da Pietro Beltrami, propriamente non rientrebbe nella stretta definizione data nel § 2.2, ma la sua importanza ed indispensabilità è tale da far passar in second'ordine ogni questione definitoria. Un'altra risorsa, accessibile ed assai curata, è il DanteSearch (http://dante.di.unipi.it:8080/DanteWeb/; cfr. Tavoni, 2011) diretto da Mirko Tavoni a Pisa: comprende tutte le opere di Dante, annotate (*Commedia*, *Convivio* e *Rime*) anche sintatticamente. La principale eccezione alla "medioevalità" pressoché esclusiva in questo tipo di corpora è, infine, costituita dall'ottocentesco CEOD (*Corpus Epistolare Ottocentesco Digitale*: http://ceod.unistrasi.it/), un corpus, coordinato da Massimo Palermo all'Università di Siena (cfr. Antonelli, Chiummo, Palermo, 2004), che raccoglie (secondo gli ultimi dati del sito) 1292 lettere, spesso inedite, di 73 scriventi, diversi per provenienza ed estrazione sociale. Interessante anche per le problematiche filologiche spesso affrontate, è completamente accessibile online.

E l'annotazione sintattica ci conduce ad un altra tipologia: il TUT (*Turin University Treebank*: http://www.di.unito.it/~tutreeb/) è un *treebank*, cioè un corpus sintatticamente annotato seguendo uno schema arborescente a dipendenza, costituito[7] da 2.860 frasi italiane e 200 inglesi[8] in allestimento da anni da parte di un altro gruppo torinese,

[7] Secondo gli ultimi dati del sito (19.04.2012).

[8] Delle cui fonti non è peraltro detto molto, anche se se ne può indurre che le parti italiane più cospicue siano tratte dal Codice civile e da generici "giornali".

centrato intorno a Leonardo Lesmo e Cristina Bosco, che hanno pubblicato diffusamente sull'argomento (cfr. ad esempio Lesmo *et alii*, 2002). Il TUT è dichiaratamente licenziato secondo *Creative Commons Share Alike*, ed è al momento largamente anche se non completamente scaricabile.

Un'ulteriore tipologia, di interesse crescente, è quella dei corpora traduzionali o interpretariali. Tra le varie iniziative attivate, quella già disponibile è legata alla SSLMIT di Forlì ed è EPIC (*European Parliament Interpreting Corpus*: http://dev.sslmit.unibo.it/corpora/corporaproject.php?path=E.P.I.C.), un corpus trilingue (italiano, inglese e spagnolo) di testi del Parlamento europeo, allineati e POS-taggati; il DIRSI-C (*DIRectionality in Simultaneous Interpreting Corpus*) di Claudio Bertazzoli (cfr. Bendazzoli, 2010), di analoga provenienza (e quindi appoggiato al medesimo staff forlivese da cui ha spiccato il volo Marco Baroni) non sembra al momento ancora disponibile.

Un altro settore in notevole espansione, un po' come l'apprendologia tutta, è quello dei corpora di apprendenti (*learner corpora*): le iniziative veramente pubbliche (come quella del citato VALICO) sono relativamente poche, ma sono la punta dell'iceberg di una pratica che è assai vasta, anche a condizioni minimali. Da ricordare, in questo genere, è senz'altro LAICO (*Lessico per Apprendere l'Italiano. Corpus di Occorrenze*: https://sites.google.com/site/corpuslaico/), coordinato a Siena da Andrea Villarini (cfr. Villarini, 2011); il corpus non è al momento interrogabile online, ma lo si può comunque fare scrivendo direttamente all'autore.

Una piccola eccezione alla italianità di origine delle risorse, come dicevamo, va fatta per l'internazionale ed assai importante (pure se parte da interessi più psicologici che linguistici) progetto CHILDES (*CHIld Language Data Exchange System*: http://childes.psy.cmu.edu/) fondato da Brian MacWhinney per studiare il linguaggio infantile, cfr. MacWhinney, 2000: tra le molte lingue in cui si articolano le sue risorse (che, peraltro, a rigore non rientrerebbero strettamente nella nostra definizione di corpus) tutte preparate in CLAN (http://childes.psy.cmu.edu/clan/) ed agevolmente scaricabili, vi è anche l'italiano (http://childes.psy.cmu.edu/data/Romance/Italian/).

Il CiT (*Corpus di Italiano Televisivo*; cfr. Spina, 2005) di Stefania Spina, infine, fino a non molto tempo fa era consultabile online (http://www.sspina.it/cit/) ma è attualmente scomparso dal Web; il che è un peccato, perché, anche se piccolo, era annotato finemente ed in modo accurato. Per fortuna ve n'è un valido successore, il LIT (*Lessico di Italiano Televisivo*: http://deckard.micc.unifi.it:8080/litsearch/), diretto da Nicoletta Maraschio ed interrogabile online. Raccoglie un campione rappresentativo dell'italiano televisivo del 2006, consistente in 168 ore di parlato tratti dalle reti RAI e Mediaset. Il Dia-LIT (http://193.205.158.203/dialit/), infine, vorrebbe estendere la campionatura del LIT all'intera storia dell'italiano televisivo, nella sua diacronia dal 1954 ad oggi. In fase di implementamento, una parte ne è già disponibile alla consultazione.

3.5. *Il parlato tra Firenze e Napoli*

L'attenzione al parlato ha una lunga tradizione in Italia, rimontando nel decennio precedente all'impresa lessicografica (lessicografia fondata su corpora nella migliore

tradizione britannica) di Tullio de Mauro del 1993. Il corpus del LIP (*Lessico di frequenza dell'Italiano Parlato*), o LIP *tout court*, che ne è derivato è attualmente consultabile sul sito BADIP di Graz (*BAnca Dati dell'Italiano Parlato*: http://badip.uni-graz.at/).

Il Corpus CLIPS (*Corpora e Lessici dell'Italiano Parlato e Scritto*: http://www.clips.unina.it/it/), creato a Napoli da Federico Albano Leoni ed interamente scaricabile previa una semplice registrazione, è basato su materiali raccolti (suddivisi tra radiotelevisivi, dialogici, letti, telefonici ed ortofonici) in 15 località italiane, oltre che "nazionali", tra il 1999 ed il 2004, presentati in veste sia audio sia testuale.

Se il CLIPS costituisce la più sicura risorsa liberamente disponibile per l'italiano parlato all'inizio del millennio, non bisogna dimenticare che anche al LABLITA (*LABoratorio Linguistico del dipartimento di ITAlianistica*: http://lablita.dit.unifi.it/) di Firenze si è lavorato lungamente sul parlato molto e bene. Il C-ORAL ROM (http://lablita.dit.unifi.it/coralrom/intro.html), che di queste ricerche è il risultato più notevole, non è tuttavia una risorsa libera (anzi è commercializzato a migliaia di euro da ELDA: http://www.elda.org/catalogue/en/speech/S0172.htm); qui la menzioniamo, oltre che per il suo intrinseco valore, perché se non pubblica è almeno "pubblicata" in quanto anche tradizionalmente edita (Cresti, Moneglia, 2005, con DVD).

Un'ultima eccezione delle eccezioni (si tratta di prodotto svizzero e "pubblicato" su CD-ROM in veste editoriale consueta) va fatta, giusta il suo intrinseco interesse, per il corpus di italiano parlato ticinese di Pandolfi, 2007.

4. IL SOFTWARE

Tra i software per la complessiva gestione e creazione di corpora scritti la posizione dominante è tenuta dal tedesco e (da alcuni anni) open source CWB (*Corpus Work Bench*: http://cwb.sourceforge.net/), inizialmente sviluppato dall'IMS Stuttgart (http://www.ims.uni-stuttgart.de): la sua eccezionale duttilità e potenza ne fanno uno strumento difficilmente sostituibile. La produzione italiana di strumenti più specifici o localizzati è tuttavia abbondante ed in genere di ottima qualità, ma prevalentemente commerciale o proprietaria: anzi, il software proprietario vanta una tradizione eccellente che va da Eugenio Picchi, autore del famoso DBT (http://www.ilc.cnr.it/pisystem/; propriamente un gestore di banche dati testuali, cfr. la definizione di corpus, *supra*) a Fabio Tamburini, autore del CORISTagger (cfr. Tamburini, 2007 e Bernardi *et alii*, 2006).

Con una eccezione importante, i *free softwares* più notevoli sono tutti legati ai gruppi di lavoro che già avevamo evidenziato.

Al gruppo torinese di b.manuel.org sono da ricondurre, per citare solo i risultati principali, la suite di strumenti (prevalentemente AWK) approntati da Marco Tomatis e e Manuel Barbera per il Corpus Taurinense (http://www.bmanuel.org/tools/C-Ttools/CTtools.html), l'analizzatore morfologico SMORFIA (*Stuttgart MORphology Finite states Italian Analyzer*: http://www.bmanuel.org/tools/SMorFIA/SMorFIA.html) ed il gestore di clitici ClitRec, entrambi di Marco Tomatis (http://www.bmanuel.org/tools/ClitRec/ClitRec.html), nonché il motore di ricerche testuali E_NT_ER (*ENgine for*

TExtual Researches: http://www.corpora.unito.it/cgi-bin/lingue/enter/enter_ index.pl?corpus=VALICO), di Adriano Allora.

Il gruppo di Marco Baroni, poi, è una vera fucina di applicazioni libere: sono almeno da ricordare il lessico di forme flesse Morph-it (http://dev.sslmit.unibo.it/linguistics/morph-it.php; cfr. Baroni, Zanchetta, 2005), validamente utilizzabile come ausilio per il tagging (cfr. *supra*,§ 2.3) dei corpora, e gli strumenti prodotti nel progetto WaCky (cfr. *supra*, § 3.2), tra cui va menzionato soprattutto il BootCat (*BOOTstrap Corpora And Terms from the web*: http://bootcat.sslmit.unibo.it/; cfr. Baroni, Bernardini, 2004) per dragare corpora dal Web.

A questi va doverosamente aggiunto il sistema GATTO di Domenico Iorio-Fili (http://www.ovi.cnr.it/index.php?page=scaricare-gatto), che è alla base della banca dati testuale di italiano antico dell'OVI (cfr. *supra*, § 3.4).

5. LA MANUALISTICA

È certo questo il campo in cui la tradizione italiana è più scarsa, non avendo ancora sviluppato trattazioni localizzate del livello e della qualità, ad esempio, di Lemnitzer, Zinsmeister, 2006 per il tedesco[9].

Una volta scartata, come fuori dai limiti di questa rassegna, la manualistica pubblicata in Italia da italiani ma in lingua inglese (ad es. Damascelli, Martelli, 2002, che si limitano peraltro a riprodurre gli schemi britannici divulgati da McEnery, Wilson, 2001; o Facchinetti, 2007, più originale) o comunque di orizzonti inglesi (ad es. Calabrese, 2004: inglese come lingua straniera; Marroni, Verzella, 2008: inglese per odontoiatri; ecc.), i saggi dedicati ad argomenti relati ma altrimenti specifici (ad es. Rovere, 2005: linguistica giuridica), o la maggior parte della dozzina di volumi miscellanei (come ad es. Moneglia, Panunzi, 2019; ma cfr. *infra*), non resta che Spina, 2002, senz'altro il manuale italiano più diffuso anche perché praticamente l'unico, di buon livello divulgativo ma ampiamente sorpassato[10], tanto più che il volume già è un rifacimento di una precedente versione del 1997. Tra l'altro non dà ancora molto conto della specifica declinazione italiana della disciplina, anche se già non era privo di spunti in questa direzione, come ad esempio l'attenzione "storica" agli albori della nostra disciplina, e la menzione dell'importanza fondante del padre Busa, di solito taciuta nei manuali britannici. Di veri manuali più recenti ci sono solo, entrambi assai utili ma per noi un poco fuori obiettivo, Lenci *et alii*, 2005, che è allargato alla linguistica computazionale in genere, e quindi riceverà la giusta attenzione in altra sezione di questo volume, e Bendazzoli 2010, che è invece ristretto ai corpora di interpretazione ed ai *Corpus-based interpreting studies* in genere. A parte ciò, per supplire all'assente manualistica, si possono consigliare solo

[9] La situazione però, sta rapidamente cambiando: già all'inizio dell'estate 2013 è uscito un manuale "tree": Barbera, 2013.

[10] Il titolo più recente in bibliografia è del 2000, ed una dozzina d'anni in questo settore rendono "fossile" qualsiasi cosa.

alcune raccolte di saggi di ampio respiro come Barbera, Corino, Onesti, 2007 e Hédiard, 2007, o al più Andorno, Rastelli, 2009, che però è limitato al solo *coté* apprendologico.

Scarso è anche il panorama delle guide web alle risorse online, di cui praticamente vi sono solo da segnalare la pagina di Isabella Chiari (http://www.alphabit.net/Corsi/IUlinks/CorporaList.htm), invero non sempre aggiornata, e quella anche più stringata del pavese LARL (Laboratorio di Analisi di Risorse Linguistiche: http://192.167.77.47/Mambo/index.php?option=com_weblinks&Itemid=23), entrambe comunque meritorie anche se assai scarne. Di ben altra dimensione e copertura era la *CLR Guide* (http://www.bmanuel.org/clr/index.html) che ho mantenuto dal 2000 fino al 2004, ma che ormai ha solo valore di documento storico, ed in quanto tale è stata lasciata online anche se i suoi links sono già quasi tutti inutilizzabili.

Un ultimo cenno va fatto alla recente voce di enciclopedia, necessariamente molto stringata, ma efficace, di Baroni, 2010.

6. RIFERIMENTI BIBLIOGRAFICI

Allora Adriano, Barbera Manuel, *Il problema legale dei corpora. Prime approssimazioni*, in Barbera, Corino, Onesti (a cura di), 2007a, pp. 109-118.

Allora Adriano, Colombo Simona, Marello Carla, *I corpora VALICO e VINCA: stranieri e italiani alle prese con le stesse attività scritte*, in Maraschio Nicoletta, De Martino Domenico, Stanchina Giulia (a cura di), *L'italiano degli altri. Atti (Firenze, 27-31 maggio 2010)*, Firenze, Accademia della Crusca, 2011, "La Piazza delle lingue" 2, pp. 49-61.

Andor József, *The Master and his Performance: An Interview with Noam Chomsky*, in «Intercultural Pragmatics», a. I, n. 1, 2004, pp. 93-111.

Andorno Cecilia, Rastelli Stefano (a cura di), *Corpora di Italiano L2: Tecnologie, metodi, spunti teorici*, Perugia, Guerra Edizioni, 2009.

Antonelli Giuseppe, Chiummo Carla, Palermo Massimo (a cura di), *La cultura epistolare nell'Ottocento. Sondaggi sulle lettere del CEOD*, Roma, Bulzoni, 2004.

Barbera Manuel, *Linguistica dei corpora italiana. Un'introduzione*, Milano, Qu.A.S.A.R., 2013; scaricabili in PDF da http://www.manuel.org/man/e-HOME.htm.

Barbera Manuel, *Per una soluzione teorica e storica dei rapporti tra grammatica generativa e linguistica dei corpora, comunicazione proposta alle 7e Giornate svizzere della linguistica. L'empiria in linguistica: varietà e complessità degli approcci*. Lugano, Università della Svizzera italiana, 13-14 settembre 2012.

Barbera Manuel, *Il neo-Corpus Taurinense e l'arte della query, comunicazione al Seminario: Sintassi dell'italiano antico e sintassi di Dante. Pisa 14-15 ottobre 2011*, ora in Tavoni Mirco, *Sintassi dell'italiano antico e sintassi di Dante*, Pisa, Felici Editore, 2012, pp. 61-79.

Barbera Manuel, *Une introduction au NUNC: histoire de la création d'un corpus*, in Ferrari, Lala, 2011, pp. 9-36.

Barbera Manuel, *Intorno a "Schema e storia del Corpus Taurinense", comunicazione al III Incontro di filologia digitale*, Verona, 3-5 marzo 2010, poi in Cotticelli Kurras Paola (a cura di), *Linguistica e filologia digitale: aspetti e progetti,* Alessandria, Edizioni Dell'Orso, 2011, pp. 27-48.

Barbera Manuel, *"Partes Orationis", "Parts of Speech", "Tagset" e dintorni. Un prospetto storico-linguistico*, in Borghi Guido, Rizza Alfredo (a cura di), *Anatolistica Indoeuropeistica e Oltre – nelle Memorie dei Seminarî offerti da Onofrio Carruba (Anni 1997-2002), al Medesimo presentate*, Milano, Qu.A.S.A.R., 2011 "Antiqui Aevi grammaticae artis studiorum consensus. Series maior" 1, tomo I, pp. 113-145.

Barbera Manuel, *Schema e storia del "Corpus Taurinense". Linguistica dei corpora dell'italiano antico*, Alessandria, Dell'Orso, 2009.

Manuel Barbera, *La resa dei forestierismi in italiano. Breve nota ortografica*, in Barbera, Corino, Onesti (a cura di), 2007 a, pp. xv-xvj.

Barbera Manuel, *Un tagset per il Corpus Taurinense. Italiano antico e linguistica dei corpora*, in Barbera, Corino, Onesti (a cura di), 2007a, pp. 135-168.

Barbera Manuel, *Mapping dei tagset in bmanuel.org / corpora.unito.it. Tra* guidelines *e prolegomeni*, in Barbera, Corino, Onesti (a cura di), 2007a, pp. 373-388.

Barbera Manuel, Corino Elisa, Onesti Cristina (a cura di), *Corpora e linguistica in rete*, Perugia, Guerra Edizioni, 2007 "L'officina della lingua. Strumenti" 1.

Barbera Manuel, Corino Elisa, Onesti Cristina, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in Barbera, Corino, Onesti (a cura di), 2007a, pp. 25-88.

Barbera Manuel, Marello Carla, *Tra scritto-parlato*, Umgangssprache *e comunicazione in rete: i corpora NUNC*, in «Studi di Grammatica Italiana», a. XXVII, 2008 (*recte* 2011) = Antonini Anna, Stefanelli Stefania (a cura di), *Per Giovanni Nencioni. Convegno Internazionale di Studi*. Pisa-Firenze, 4-5 Maggio 2009, Firenze, Le Lettere, 2011, pp. 157-185.

Barbera Manuel, Onesti Cristina, *Dalla Valsusa in avanti: i corpora di stampa periodica locale*, in «Rivista Internazionale di Tecnica della Traduzione | International Journal of Translation», a. XII, 2010, Special issue *Traduzioni nella stampa periodica*, a cura di Ondelli Stefano, pp. 103-116.

Baroni Marco, *Corpora di lingua italiana*, in Simone Raffaele (a cura di), *Enciclopedia dell'italiano*, Roma, Istituto dell'Enciclopedia italiana (fondata da Giovanni Treccani), vol. I, 2010, "Vocabolario Treccani", *sub vocem* (pp. 300b-303a).

Baroni Marco, Bernardini Silvia (a cura di), *WaCky! Working Papers on the Web as Corpus*, Bologna, GEDIT edizioni, 2006, disponibile online alla pagina http:// wackybook.sslmit.unibo.it/

Baroni Marco, Bernardini Silvia, Ferraresi Adriano, Zanchetta Eros, *The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-crawled Corpora*, in «Journal of Language Resources and Evaluation», a. XLIII, n. 3, 2009, pp. 209-226.

Baroni Marco, Bernardini Silvia, Comastri Federica, Piccioni Lorenzo, Volpi Alessandra, Aston Guy, Mazzoleni Marco, *Introducing the* La Repubblica *Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian*, in *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, ELDA, 2004, pp. 1771-1774, online alla pagina http://www.form. unitn.it/~baroni/publications/lrec2004/rep_lrec_2004.pdf.

Baroni Marco, Evert Stefan, *Statistical Methods for Corpus Exploitation*, in Lüdeling Anke, Kytö Merja (a cura di), *Corpus Linguistics, An International Handbook*, Volume 2, Berlin, Mouton de Gruyter, 2009, pp. 777-802.

Baroni Marco, Zanchetta Eros, *Morph-it! A free corpus-based morphological resource for the Italian language*, in *Proceedings of Corpus Linguistics 2005*, Birmingham, University of Birmingham, 2005; online alla pagina http://home.sslmit.unibo.it/eros/downloads/Morph-it.pdf

Bendazzoli Claudio, *Corpora e interpretazione simultanea*, Bologna, Asterisco, 2010.

Bernardi Raffaella, Bolognesi Andrea, *Seidenari Corrado, Tamburini Fabio, POS Tagset Design for Italian*, in *Proceedings of the 5th International Conference on Language Resources and Evaluation-LREC 2006*, Genova, 2006, pp. 1396-1401.

Biber Douglas, *Using Computer-based Text Corpora to Analize the Referential Strategies of Spoken and Written Texts*, in Svartvik (a cura di), 1992, pp. 213-252.

Biber Douglas, Finegan Edward, *On the Exploitation of Computerized Corpora in Variation Studies*, in Aijmer Karin, Altenberg Bengt (a cura di), *English Corpus Linguistics. Studies in Honor of Jan Svartvik*, London-New York, Longman, 1991, pp. 204-220.

Borghetti Claudia, Castagnoli Sara, Brunello Marco, *I testi del web: una proposta di classificazione sulla base del corpus PAISÀ*, in Cerruti Massimo, Corino Elisa, Onesti Cristina (a cura di), *Formale e informale. La variazione di registro nella comunicazione elettronica*, Roma, Carocci Editore, 2011, "Biblioteca di testi e studi" 683, pp. 147-170.

Burnard Lou, Bauman Syd (a cura di), *TEI P5:Guidelines for Electronic Text Encoding and Interchange*, Charlottesville (Virginia), Text Encoding Initiative Consortium, 2011; online alla pagina http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TitlePageVerso.html.

Burr Elisabeth, *Das Korpus romanischer Zeitungssprachen in Forschung und Lehre*, in Dahmen Wolfgang, Holtus Günter, Kramer Johannes, Metzeltin Michael, Schweickard Wolfgang, Winkelmann Otto (a cura di), *Romanistik und neue Medien. Romanistisches Kolloquim XVI*, Tübingen, Günther Narr, 2004 "TBL" 4555, pp. 133-62.

Buzzetti Dino, *Rappresentazione digitale e modello del testo*, in *Il ruolo del modello nella scienza e nel sapere. Roma, 27-28 ottobre 1998*, Roma, Accademia Nazionale dei Lincei, 1999, "Contributi del Centro Linceo Interdisciplinare 'Beniamino Segre'" 100, pp. 127-161.

Calabrese Rita, *La linguistica dei corpora e l'inglese come lingua straniera*, Napoli, Massa, 2004.

Carpenter Bob, *The Logic of Typed Feature Structures. With Application to Unification Grammars, Logic Programs, and Constraint Resolution*, Cambridge (UK), Cambridge University Press, 1992 "Cambridge Tracts in Theoretical Computer Science" 32.

Ciurcina Marco, Ricolfi Marco, *Le Creative Commons Public Licences per i corpora. Una suite di modelli per la linguistica dei corpora*, in Barbera, Corino, Onesti (a cura di), 2007a, pp. 127-132.

Corino Elisa, Marello Carla (a cura di), *VALICO. Studi di linguistica e didattica*, Perugia, Guerra, 2009.

Cresti Emanuela, Moneglia Massimo (a cura di), *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam-Philadelphia, John Benjamins Publishing Company, 2005, "Studies in Corpus Linguistics" 15.

Damascelli Adriana Teresa, Martelli Aurelia, *Corpus Linguistics and Computational Linguistics: an Overview with Special Reference to English*, Torino, Celid, 2002.

Facchinetti Roberta, *Theoretical Description and Practical Applications of Linguistic Corpora*, Verona, QuiEdit, 2007.

Ferrari Angela, Lala Letizia (a cura di), *Variétés syntaxiques dans la variété des textes online en italien: aspects micro- et macrostructuraux*, Nancy, Université de Nancy II, 2011 = «Verbum», a. XIII, nn. 1-2, 2011.

Fillmore Charles J., *"Corpus Linguistics" or "Computer-aided Armchair Linguistics"*, in Svartvik (a cura di), 1992, pp. 35-60.

Hédiard Marie (a cura di), *Linguistica dei corpora. Strumenti e applicazioni*, Cassino, Edizioni dell'Università degli studi di Cassino, 2007, "Collana Scientifica" 20.

Kilgarriff Adam, Grefenstette Gregory, *Introduction to the Special Issue on the Web as Corpus*, in «Computational Linguistics», a. XXIX, n. 3, 2003, pp. 333-347, disponibile anche online alla pagina http://www.kilgarriff.co.uk/publications.htm.

Lenci Alessandro, Montemagni Simonetta, Pirrelli Vito, *Testo e computer. Elementi di linguistica computazionale*, Roma, Carocci, 2005, "Università" 664.

Lemnitzer Lothar, Zinsmeister Heike, *Korpuslinguistik: eine Einführung*, Tübingen, Gunter Narr Verlag, 2006 "Narr Studienbücher".

Lesmo Leonardo, Lombardo Vincenzo, Bosco Cristina, *Treebank Development: the TUT Approach, in Proceedings of the International Conference on Natural Language Processing (ICON 2002)*, Mumbay, India, 2002; online a http://www.di.unito.it/~tutreeb/.

MacWhinney Brian, *The CHILDES Project: Tools for Analyzing Talk*. Volume 1: *Transcription format and programs*. Volume 2: *The Database*, Mahwah (NJ), Lawrence Erlbaum Associates, 2000.

Manning Christopher D., Schütze Hinrich, *Foundations of Statistical Natural Language Processing*, Cambridge (Massachusetts)-London (England), The MIT Press, 2000^3 [1999_1].

Marello Carla, *Le parole dell'italiano. Lessico e dizionari,* Bologna, Zanichelli, 1996.

Marroni Michela, Verzella Massimo, *Linguistica dei corpora. Inglese specialistico e odontoiatria*, Roma, Aracne, 2008, "Studi di anglistica".

McEnery Tony, Wilson Andrew, *Corpus Linguistics. An Introduction*, Edinburgh, Edinburgh University Press, 2001_2 [1996_1, 2005^r] "Edinburgh Textbooks in Empirical Linguistics".

Monachini Monica, *ELM-IT: EAGLES Specifications for Italian Morphosyntax-Lexicon Specifications and Classification Guidelines*, Pisa, EAGLES Document EAG-CLWG-ELM-IT/F, May 1996.

Moneglia Massimo, Panunzi Alessandro (a cura di), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*, Firenze, Firenze University Press, 2010.

Onesti Cristina, *Methodology for Building a Text-Structure Oriented Legal Corpus*, in «Comparative Legilinguistics» a. VIII, 2011, pp. 37-50.

Pandolfi Elena Maria, *Misurare la regionalità. Uno studio quantitativo su regionalismi e forestierismi nell'italiano parlato nel Canton Ticino*, Locarno, Osservatorio linguistico della Svizzera italiana-Armando Dadò, 2006.

Peirce Charles Sanders, *Prolegomena to an Apology for Pragmaticism*, 1906, in Hartshorne Charles, Weiss Paul (a cura di), *Collected Papers of Charles Sanders Peirce*, vol. III *Exact Logic (Published Papers)* and vol. IV *The Simplest Mathematics*, Cambridge (Mass.), Harvard University Press, 1933, p. 537.

Quine Willard van Orman, *Quiddities: an Intermittently Philosophical Dictionary*, Cambridge (Mass.), the Belknap Press of Harvard University Press, 1987.

Ramsay Allan, *Discourse*, in Mitkov Ruslan (a cura di), *The Oxford Handbook of Computational Linguistics*, Oxford-New York, Oxford University Press, 2003, pp. 112-135.

Renzi Lorenzo, *Il progetto ItalAnt e la "grammatica del corpus"*, in *Lingue romanze nel Medioevo. Atti del convegno, Piliscsaba, 22-23 marzo 2002*, Domokos Györgyi, Salvi Giampaolo (a cura di) = «Verbum. Analecta Neolatina», a. IV, n. 2, 2002, pp. 271-94.

Renzi Lorenzo, *L'autobiografia linguistica in generale, e quella dell'autore in particolare, con un saggio di quest'ultima*, in Cini Monica, Regis Riccardo (a cura di), *Che cosa ne pensa oggi Chiaffredo Roux? Percorsi di dialettologia percezionale all'alba del nuovo millennio. Atti del Convegno internazionale (Bardonecchia, 25-27 maggio 2000)*, Alessandria, Edizioni Dell'Orso, 2002, pp. 329-339, poi in Renzi Lorenzo, *Le piccole strutture. Linguistica, poetica, letteratura*, a cura di Andreose Alvise, Barbieri Alvaro, Cepraga Dan Octavian, Bologna, Società editrice il Mulino, 2008, pp. 3-16.

Renzi Lorenzo, Salvi Giampaolo (a cura di), *Grammatica dell'italiano antico*, 2 voll., Bologna, il Mulino, 2010.

Rovere Giovanni, *Capitoli di linguistica giuridica. Ricerche su corpora elettronici*, Alessandria, Edizioni Dell'Orso, 2005, "Gli strumenti umani" 9.

Sabatini Francesco, *La storia dell'italiano nella prospettiva della corpus linguistics*, in Corino Elisa, Marello Carla, Onesti Cristina (a cura di), *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 | Proceedings of the XII EURALEX International Congress*. Torino, Italia, 6th-9th September 2006, 2 voll., Alessandria, Edizioni Dell'Orso, 2006, pp. 31-37.

Sampson Geoffrey, *Empirical Linguistics*, London-New York, Continuum, 2001.

Sinclair John [McHardy], *Corpus, Concordance, Collocation*, Oxford, Oxford University Press, 1991.

Spina Stefania, *Fare i conti con le parole. Introduzione alla linguistica dei corpora*, Perugia, Guerra Edizioni, 2001.

Spina Stefania, *Il Corpus di italiano televisivo (CiT): struttura e annotazione*, in Burr Elisabeth (a cura di), *Tradizione & Innovazione. Il parlato: teoria-corpora-linguistica dei corpora. Atti del VI Convegno SILFI (Gerhard Mercator Universität Duisburg 28 giugno-2 luglio 2000)*, Firenze, Franco Cesati Editore, 2005, pp. 413-426.

Svartvik Jan (a cura di), *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82. Stockholm, 4-8 August 1991*, Berlin, Mouton de Gruyter, 1992, "Trends in Linguistics. Studies and Monographs" 65.

Tamburini Fabio, *CORISTagger: a High Performance PoS Tagger for Italian*, in «Intelligenza artificiale», a. IV, n. 2, 2007, pp. 14-15.

Tavoni Mirko, *DanteSearch: il corpus delle opere volgari e latine di Dante lemmatizzate con marcatura grammaticale e sintattica*, in Cerbo Anna, di Fiore Ciro (a cura di), *Lectura Dantis in onore di Vincenzo Placella*, Napoli, Liguori, 2011, pp. 567-591.

Villarini Andrea, *La competenza lessicale: un viaggio tra libri di testo e parlato del docente*, in Jafracesco Elisabetta (a cura di), *L'acquisizione del lessico nell'apprendimento*

dell'italiano L2. Atti del XIX convegno nazionale ILSA, Firenze, 27 novembre 2010, Firenze, Le Monnier, 2011, "Italiano per stranieri", pp. 53-80.

Zanni Samantha, *Corpora elettronici e copyright. Lo stato legale della questione*, in Barbera, Corino, Onesti (a cura di), 2007a, pp. 119-126.